ACS OMEGA

Article

Cite This: ACS Omega 2019, 4, 6421−6435

http://pubs.acs.org/journal/acsodf

Analysis of Gene Expression Differences between Different Pancreatic Cells

Lei Chen,[†,‡,§] Xiaoyong Pan,[∥] Yu-Hang Zhang,[⊥] Tao Huang,*[,⊥] and Yu-Dong Cai*[,†]

[†]School of Life Sciences, Shanghai University, Shanghai 200444, China

[‡]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China

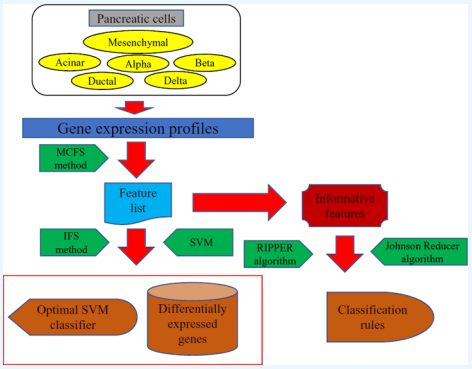
[§]Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China

[∥]Department of Medical Informatics, Erasmus MC, Rotterdam 3014ZK, Netherlands

[⊥]Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Supporting Information

ABSTRACT: The pancreas is a complex glandular organ in the abdominal cavity of humans. Pancreatic tissues contain different function regions that participate in different biological processes. Moreover, pancreatic cells can be clustered into different types according to their location regions. Cells in different types are involved in different biological functions as either endocrine or digestive. In this study, we investigated the gene expression of pancreatic cells in six types, tried to identify differentially expressed genes among different cell types, and obtained pancreatic cell biomarkers. In detail, the Monte Carlo feature selection (MCFS) method was performed on the dataset consisting of the gene expression of 2282 pancreatic cells. The resulting feature list was further used in the incremental feature selection method, with the help of a support vector machine to extract important differentially expressed genes. In addition, on the basis of the 776 informative features yielded by the MCFS method, we set up 12 classification rules via Johnson Reducer algorithm and Repeated Incremental Pruning to Produce Error Reduction algorithm, which can assign cells into six types. We did the enrichment analysis on obtained differentially expressed genes and extensively analyzed the top 10 differentially expressed genes and rules via literature reviewing, indicating that the results of this study are quite reliable.



1. INTRODUCTION

The pancreas, a complicated glandular organ in the abdominal cavity of humans, is a multifunctional organ that contributes to both digestive and endocrine systems.[1,2] Physically, the pancreas is located in the upper belly behind the stomach and in front of the spine and can be further divided into four major subregions: (1) head, (2) neck, (3) body, and (4) tail of pancreas, which have different organization structures, biological functions, and anatomical positions.[2]

For the complicated biological functions of the pancreas in the endocrine and exocrine systems, the two major biological functions that the pancreas participates in include endocrine blood glucose regulation and exocrine pancreatic juice secretion.[3−5] Regulated by a negative feedback mechanism, the general blood glucose levels are accurately regulated by two hormones (insulin and glucagon) secreted by the pancreatic islets of the pancreas.[6] For its digestive function, the pancreas secretes pancreatic juice containing multiple enzymes into the duodenum and contributes to further digestion of intestinal food.[7] To maintain the general function of these two major biological processes, two independent systems have been built in the pancreatic tissues that are involved in different cell subgroups and potential mechanisms.[6] Three subgroups of cells that participate in the regulation of sugar are located in the pancreas: (1) cells located in pancreatic islets, including alpha, beta, pancreatic polypeptide (PP), and gamma cells, which directly secrete functional hormones, such as insulin and glucagon;[8,9] (2) sympathetic cells, including alpha 2 and beta 2 cells, which contribute to the direct regulation of alpha and beta cells;[10] and (3) parasympathetic cells, including M3 cells, which also regulate the stimulation of alpha and beta cells.[11] These cells contribute to the secretion of specific hormones or their regulators, making them quite different from digestive system-associated cells. Digestion-associated biological processes are regulated by functional cells filled with granules containing effective enzymes located all over the pancreatic tissues, except on the endocrine regions.[12]

As mentioned above, pancreatic tissues contain different function regions participating in different biological processes. Cells located in different regions can be clustered into different subgroups, which are involved in different biological functions as either endocrinal or digestive, like the alpha cells in the

Received: August 26, 2018
Accepted: February 26, 2019
Published: April 8, 2019

pancreatic islets. Therefore, considering the complicated biological functions of the pancreas, its cellular component may be complicated and diverse. Including the stroma and immune cells in the microenvironment of the pancreas, more than 10 subgroups of cells are present in this multifunctional organ.[13,14] Each cell subgroup has biological functions mediated by their respective protein distribution and gene expression patterns, which may quite vary from the others. Therefore, protein distribution and gene expression patterns may be good markers for the distinction of different cell subgroups. However, traditionally, it is quite hard to recognize the gene expression profile at the single-cell level, making cell clustering based on gene expression patterns hard. With the development of single-cell RNA sequencing technologies, the identification of detailed gene expression profiling at the single-cell level has been achieved,[15−18] making detailed cell clustering at the single-cell level possible based on its transcriptomic characteristics.

In this study, on the basis of single-cell RNA sequencing technologies, we tried to recognize the expression profiling of different cell types in pancreatic tissues at the single-cell level. According to the histological and morphological characteristics of pancreatic cells, we artificially clustered all candidate cells into six optimal subgroups: (1) acinar, (2) alpha, (3) beta, (4) delta, (5) ductal, and (6) mesenchymal cells. Acinar cells, also known as spindle-shaped duct cells in the exocrine pancreas, have been widely reported to participate in the formation of the main pancreatic duct and secrete an aqueous bicarbonate solution under hormone stimulation.[19] Pancreatic alpha cells, as another functional cell subgroup, have been widely reported to synthesize and secrete the peptide hormone glucagon, regulating the glucose levels in blood.[20] As for beta cells, such pancreatic cancer subgroup participate in the storage and release of insulin, making up 65−80% of the pancreatic islets.[21] The delta cells in pancreatic tissues have been widely regarded as somatostatin-producing cells suppressing the release of multiple pancreatic hormones including insulin.[22] As for the ductal cells in pancreatic ducts, different from spindle-shaped acinar cells, such subgroup of cells lines the pancreatic duct and participates in the maintenance of pH in the pancreatic duct.[23] Mesenchymal cells refer to all the stroma cells in the pancreatic tissues, participating in the regulation of regional immune response and the maintenance of the pancreatic microenvironment.[24,25]

On the basis of the detailed expression profiling data at the single-cell level provided by Enge et al.,[26] we applied several machine learning algorithms to identify potential quantitative and qualitative biomarkers/standards, which can be used for the distinction of six optimal subgroups of cells. Compared with previous statistical approaches reviewed by Soneson et al.,[27] the methods used here considered the complex relationships among genes and optimized the gene selection. In detail, previous statistical approaches only evaluated the associations between genes and samples, whereas the machine learning algorithms can further consider the relationships among genes. It is believed that we can find more compact signatures, that is, few genes with the same or better performance, which were more suitable for biomarkers. The procedures first applied the Monte Carlo feature selection (MCFS)[28] method to analyze expression profiling data, producing a ranked feature list according to importance. Then, the incremental feature selection (IFS)[29] method, together with the support vector machine (SVM),[30] was employed to extract optimal features that yielded the best performance for SVM to the distinction of six optimal subgroups of cells. The corresponding genes of optimal features were accessed, which can be potential biomarkers. We did the enrichment analysis on these optimal genes and extensively discussed the top 10 genes to prove the reliability of our results. On the other hand, the MCFS method produced 776 informative features, on which 12 rules were constructed via Johnson Reducer algorithm[31] and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm[32] to classify cells. These rules can give a clearer outline of cells in six groups and were further extensively analyzed. This study not only can contribute to the identification of potential biomarkers for each cell subgroup but can also help reveal cell subgroup-specific biological processes, drawing a panorama for the complicated tissue components in pancreatic tissues.

2. RESULTS

In this study, we tried to analyze the gene expression profiles of six subgroups of cells of the human pancreas. To do that, several advanced computational methods, including MCFS, IFS, SVM, RIPPER, and Johnson Reducer algorithm, were employed. The whole procedures are illustrated in Figure 1.

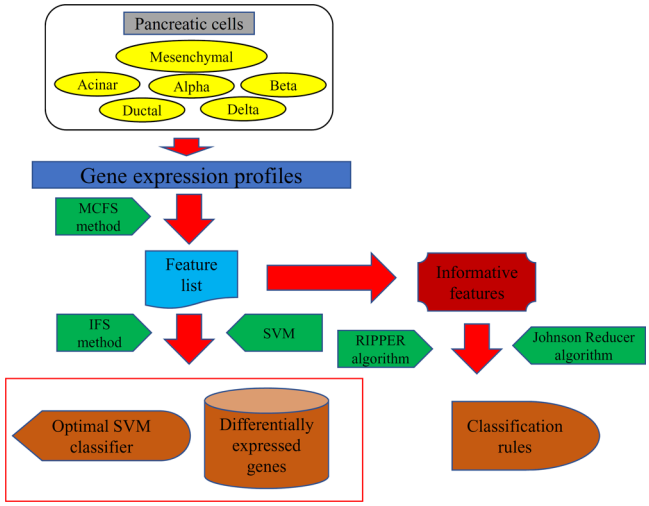


Figure 1. Procedures of investigating pancreatic cells in six subgroups via several machine learning algorithms. First, all pancreatic cells were represented by the expression levels of 23,459 genes. Second, the Monte Carlo feature selection (MCFS) method was adopted to analyze these features (genes), resulting in a feature list. Then, for the obtained feature list, on the one hand, we employed the incremental feature selection (IFS) method and support vector machine (SVM) to extract differentially expressed genes and construct an optimal SVM classifier; on the other hand, the informative features (some top features in the list) were picked up to produce classification rules via Johnson Reducer algorithm and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm.

2.1. Results of MCFS Method. According to Figure 1, the MCFS method was applied on the gene expression profiles of pancreatic cells in six subgroups. The importance of each feature (gene) was evaluated by an RI score. Then, a feature list *F* was obtained, in which all features were ranked according to descending RI scores, which are provided in Table S1.

2.2. Results of IFS Method with SVM. In the feature list *F*, the rank of each feature can indicate its importance. However, it is difficult to determine which features can be

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

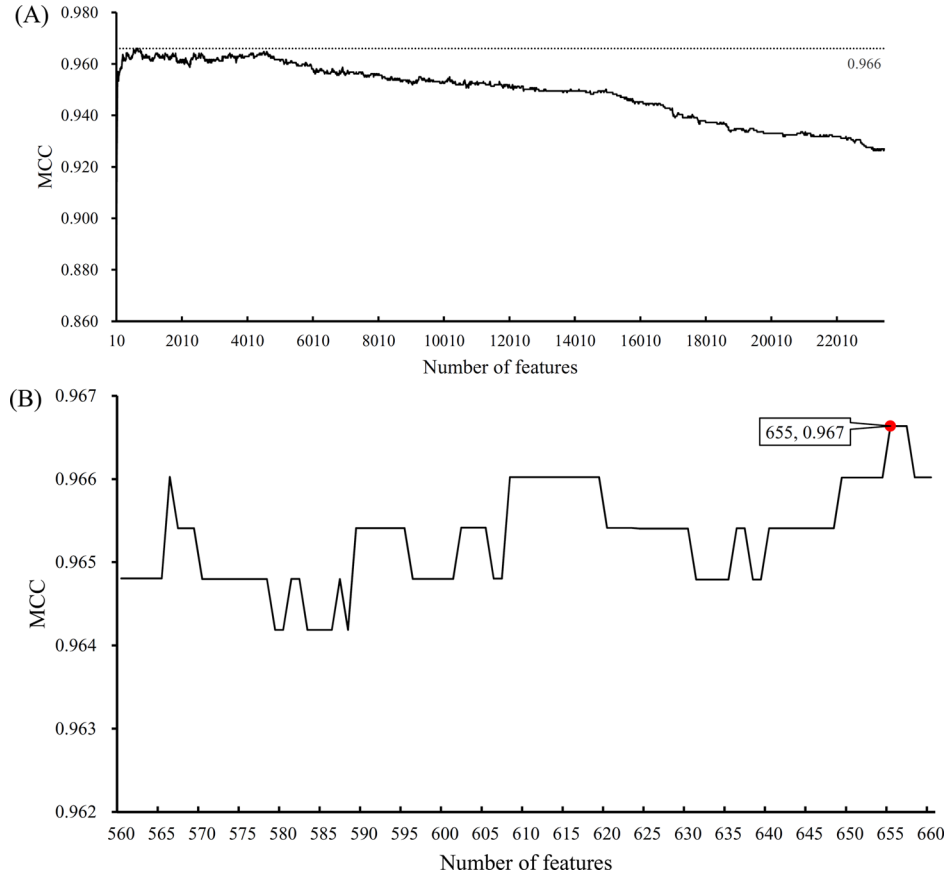


Figure 2. IFS curves to illustrate the classification performance of SVM classifiers by using different feature subsets. The *x* axis represents the number of features participating in the classification, and the *y* axis represents the MCC value. (A) IFS curve based on the top multiple of 10 features in the list yielded by the MCFS method. High MCC values (no less than 0.966) were gathered in the interval [560, 660]. (B) IFS curve based on the feature subsets consisting of the top 560−660 features in the feature list yielded by the MCFS method. Using the top 655 features, the highest MCC of 0.967 was accessed.

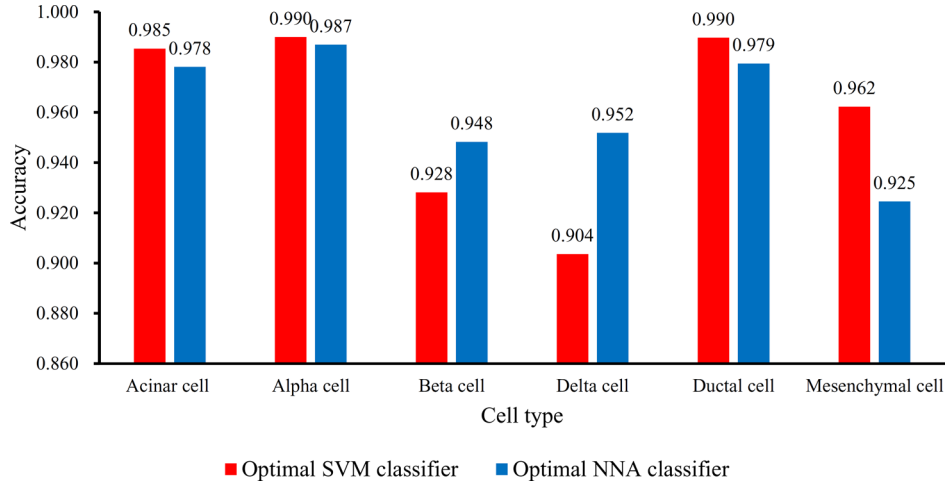


Figure 3. Bar chart to illustrate the performance of the optimal SVM/NNA classifier on each cell subgroup. The optimal SVM classifier yielded six accuracies higher than 0.9, indicating the effectiveness of this classifier. Also, it is a bit superior to the optimal NNA classifier.

optimal for a given classification algorithm. To this end, the IFS method combined with a multiclass SVM was employed, which would be introduced in Sections 4.2 and 4.3. To save time, the IFS method was divided into two stages. In the first stage, the IFS method was used to construct a series of feature subsets with step 10; that, F_{10}, F_{20},...,F_{23450} were constructed, where F_i contained top i features in F, and i was a multiple of 10. Then, SVM was adopted to test each feature subset; that is, SVM was executed on all cells that are represented by features in the feature subset. The predicted results were counted as accuracies for each type, ACC and MCC, which are available in Table S2. For easy observation, we plotted a curve in Figure 2A, in which the *x* axis represents the number of features participating in the classification and *y* axis represents the

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

Table 1. MCC of Each Subgroup Yielded by the Optimal (Support Machine Vector) SVM Classifier and Nearest Neighbor Algorithm (NNA) Classifier

classification algorithm	acinar cell	alpha cell	beta cell	delta cell	ductal cell	mesenchymal cell
SVM	0.979	0.956	0.954	0.948	0.986	0.981
NNA	0.976	0.966	0.956	0.974	0.964	0.951

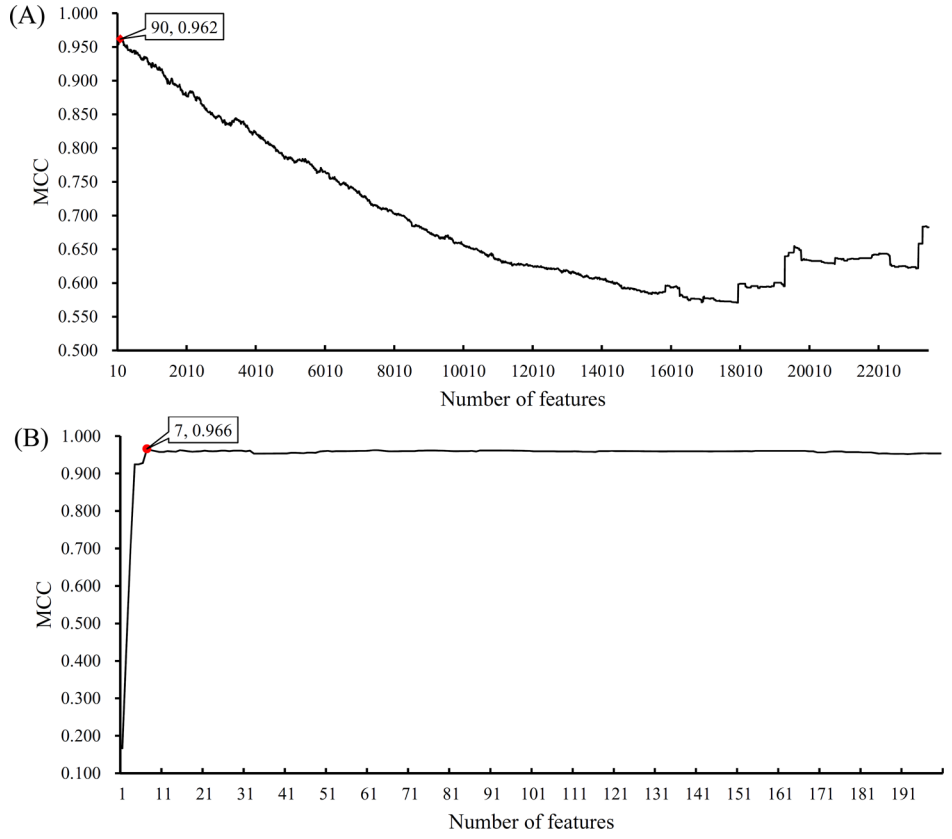


Figure 4. IFS curves to illustrate the classification performance of NNA classifiers by using different feature subsets. The *x* axis represents the number of features participating in the classification, and the *y* axis represents the MCC value. (A) IFS curve based on the top multiple of 10 features in the feature list. The highest MCC is 0.962 when the top 90 features are used. We determine the interval as [1, 200]. (B) IFS curve based on the feature subsets consisting of the top 1−200 features in the feature list. Using the top seven features, the highest MCC of 0.966 was obtained.

MCC. This curve first follows a sharp increasing trend and then follows a decreasing trend upon reaching its maximum. The maximum MCC was 0.966 when the top 610, 660, or 650 features were used to represent cells, which were the only points to reach 0.966. To further determine the number of features in the optimal feature set, we yielded a number interval [560, 660], in which a feature set may yield a better performance. Thus, in the second stage, we tested all possible feature subsets in this interval. In detail, we generated a series of feature subsets with step 1 within the interval [560, 660], and SVM was performed on all cells represented by the features in each subset. The predicted results were also counted as accuracies for six types, ACC and MCC, which are listed in Table S3. To clearly exhibit the performance of SVM on these feature subsets, a curve was plotted in Figure 2B, which was defined as the same as that in Figure 2A. Clearly, when the top 655 features in F were used, the MCC first reached a maximum of 0.967. Accordingly, these features were called optimal features for SVM and comprised the optimal feature set. The SVM constructed based on them was the optimal SVM classifier. The performance of the optimal SVM classifier on each cell type is illustrated in Figure 3. The accuracies for all types were all higher than 0.9, and the ACC was 0.976. In addition, we also computed the MCC for each subgroup, listed in Table 1, from which we can see that each MCC was higher than 0.940. All of these indicate the good performance of this classifier. The largest subgroup contained 998 cells, and the smallest one consisted of 53 cells, meaning that the investigated dataset was imbalanced. Generally, the predicted results would be apt to subgroups with large sizes. However, according the accuracies in Figure 3 and MCCs in Table 1, they were all quite high (larger than 0.9), suggesting that this fact did not affect the results a lot in this study.

2.3. Comparison of IFS Method with NNA. In this study, we selected the SVM as the classification algorithm and constructed an optimal SVM classifier with good performance (MCC = 0.967). Here, we employed another classification algorithm, nearest neighbor algorithm (NNA), for comparison. The same procedures for SVM were applied to NNA. First, for each of feature sets F_{10}, F_{20},...,F_{23450}, the NNA was executed on all pancreatic cells represented by features in each of these sets. The results were counted as accuracies for each subgroup, ACC and MCC, which are available in Table S4. Similarly, we plotted an IFS curve in Figure 4A with the same settings in

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

Table 2. Twelve Classification Rules Yielded by Johnson Reducer Algorithm for Classifying Cells into Different Subgroups

rule tag	condition	outcome	support[a]	accuracy[b]
rule-1	*SPARCL1* ≥ −0.075 *CRISPLD2* ≥ −0.089	mesenchymal cell	0.023	1.000
rule-2	*SST* ≥ 0.363	delta cell	0.037	0.940
rule-3	*SST* ≥ 0.126	delta cell	0.038	0.943
rule-4	*INS* ≥ −0.220	beta cell	0.156	0.946
rule-5	*INS* ≥ −0.357 *CALM1* ≤ −0.601 *CADPS* ≥ −0.258	beta cell	0.017	0.921
rule-6	*CFTR* ≥ −0.305 *ANXA4* ≥ 0.136	ductal cell	0.145	0.994
rule-7	*TFPI2* ≥ −0.216 *SPINK1* ≤ −0.349 *CHGB* ≤ −0.491	ductal cell	0.131	0.930
rule-8	*TINAGL1* ≥ 0.617	ductal cell	0.067	0.928
rule-9	*PROM1* ≥ −0.174 *ALDH1A1* ≤ −0.595	ductal cell	0.026	0.950
rule-10	*TRY6* ≥ −0.295	acinar cell	0.189	0.940
rule-11	*LGALS3* ≥ −0.204	acinar cell	0.226	0.298
rule-12	other conditions	alpha cell	0.433	0.992

[a]Support is defined as the proportion of all samples satisfying the condition. [b]Accuracy is defined as the proportion of the corrected classified samples among those satisfying the condition.

Figure 2A. It can be observed that the maximum MCC was 0.962 when the top 90 features were used. Thus, we determined the interval as [1, 200] to perform the second stage of the IFS method. The predicted results are also provided in Table S4, and the IFS curve is plotted in Figure 4B. The highest MCC was 0.966 when the top seven features were used to construct the classifier. Accordingly, for NNA, these seven features were optimal features based on which optimal NNA classifier was built. The detailed performance, including accuracies for six subgroups, of such optimal NNA classifier is shown in Figure 3, and the MCC for each subgroup is listed in Table 1.

The optimal NNA classifier yielded the MCC of 0.966, which was a bit lower than that of the optimal SVM classifier. For six accuracies of the six subgroups (Figure 3), the optimal SVM classifier provided higher values on four subgroups, and the optimal NNA classifier yielded higher accuracies on the other two subgroups. For the MCCs of six subgroups (Table 1), each optimal classifier defeated the other one on three subgroups. Considering the fact that the optimal SVM classifier yielded a higher MCC (in multiclass), we considered that the optimal SVM classifier was a bit superior to the optimal NNA classifier.

2.4. Results of Rule Learning. As mentioned above, the optimal SVM classifier can yield good performance for classifying pancreatic cells into six subgroups. However, it was totally a black box. It is quite difficult to uncover its classification procedures, thereby giving limited insights for understanding the gene expression differences of pancreatic cells in different subgroups. Thus, the Johnson Reducer algorithm and RIPPER algorithm were adopted to extract classification rules using the 776 informative features yielded by the MCFS method. Here, we obtained 12 classification rules, listed in Table 2, via above-mentioned algorithms that are integrated in the program of the MCFS method used in this study. We also calculated the support and accuracy of each rule, which are also listed in Table 2. All rules except rule-11 received the accuracies higher than 0.9, indicating the utility of these rules. In addition, to evaluate the performance of the rules yielded by Johnson Reducer algorithm and RIPPER algorithm on informative features, the 10-fold cross-validation was executed thrice, yielding an ACC of 0.965. The confusion map is shown in Figure 5. We also counted the MCC (multiclass version) of the predicted results, yielding an MCC of 0.923. Although it is lower than that obtained by the optimal SVM classifier, it can give a clear procedure of classification and provide a clearer outline of the differences between cells in different subgroups.

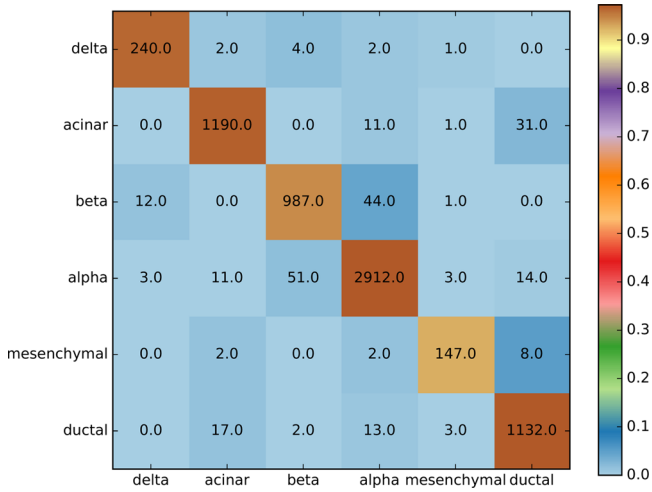


Figure 5. Confusion matrix for 10-fold cross-validation by using the 12 classification rules listed in Table 2 for classifying cells into six subgroups. The numbers were pooled from running 10-fold cross-validation thrice. The row represents the actual cell subgroup, and the column indicates the predicted cell subgroup.

In addition, to further test the utility of above-mentioned rules, we employed an independent test dataset that was retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73727.[33,34] In this dataset, mesenchymal cells were not included. There were 11 acinar cells, 18 alpha cells, 12 beta cells, 2 delta cells, and 8 ductal cells. The whole dataset is provided in Supplementary Data S1. Because the gene *TRY6* was not measured in this dataset and one cell type (mesenchymal cells) was not included, we discarded rule-1 (for mesenchymal cells) and rule-10 (using *TRY6*) and used the remaining 10 rules to predict the cell type of each cell in such dataset. As a result, 47 cells were correctly predicted, inducing the ACC of 92.16%. In detail, the accuracies for five cell types were 63.64, 100, 100, 100, and 100%, respectively. The MCC was 0.901. All these results indicate that the constructed rules also provided good performance on such independent test dataset, implying the good generalization of these rules.

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

3. DISCUSSION

In this study, we applied several advanced computational methods to analyze the gene expression profile of pancreatic cells in six subgroups. Several optimal features (genes) and an optimal classifier were obtained. In addition, 12 classification rules were generated. In this section, some of the obtained genes and classification rules were extensively analyzed and discussed.

3.1. Analysis of Optimal Differential Expression Genes. A total of 655 features were used to construct the optimal SVM classifier. To test the statistical significance of these features, we produced 1000 feature sets. Each set contained 655 features that were randomly selected from all features. For each set, an SVM classifier was built, and its performance was evaluated with 10-fold cross-validation. Accordingly, we accessed 1000 MCCs (multiclass version). A box plot was drawn in Figure 6 to show these MCCs. The

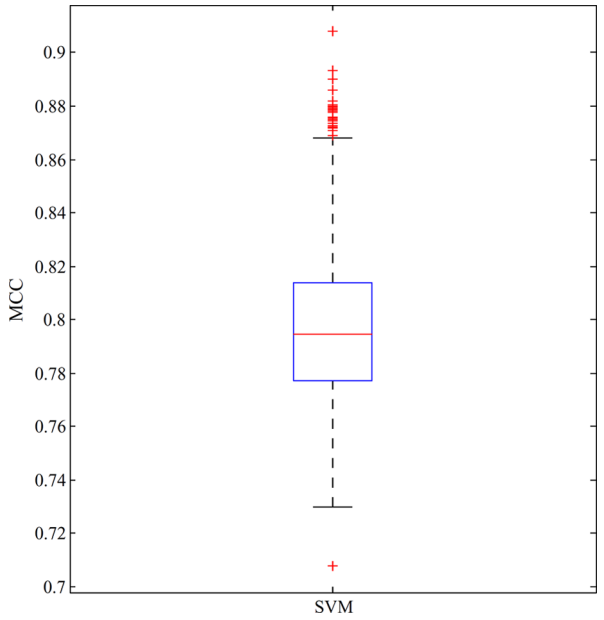


Figure 6. Box plot to illustrate the distribution of MCCs yielded by SVM classifiers on 1000 randomly produced feature sets. Each feature set contained 655 features. The optimal SVM classifier built on the top 655 features produced the MCC of 0.967. Clearly, these 655 features are statistically significant because they can generate the MCC much higher than MCCs shown in the box plot.

MCC produced by the optimal SVM classifier was 0.967. It can be observed from Figure 6 that such value was much higher than 1000 MCCs on randomly produced sets. In addition, the mean and standard deviation of 1000 MCCs were 0.797 and 0.030, respectively, indicating that 0.967 was higher than the mean plus 5.6 standard deviations. It is suggested that 655 optimal features for SVM were statistically significant. It was helpful to uncover gene expression differences between cells in six subgroups by analyzing these features.

3.1.1. Enrichment Analysis on Optimal Features. Before detailed analysis of some optimal features (genes), we first presented a gene ontology (GO) and KEGG pathway enrichment analysis on 655 optimal genes (Table S5). These optimal genes may represent the specific expression pattern of their respective cell subgroups and probably contribute to the cell-specific biological processes. Here, we selected typical GO terms or KEGG pathways for detailed discussion, which reflected the specific biological functions of one or more cell subgroups.

A specific GO, describing the molecular function of glycosaminoglycan binding (GO: 0005539) with a false discovery rate (FDR) of 9.74×10^{-9}, was enriched by optimal genes. As we all know, according to recent publications,[35,36] glycosaminoglycan binding has been reported to participate in multiple physical biological processes, and the abnormal regulation of such molecular function may result in malignant transformation of pancreatic cells.[37] Therefore, the regulation of such molecular function may be quite significant for pancreatic physical metabolism, and it is quite reasonable to enrich the optimal genes into such molecular function. Apart from that, the top six GO terms (GO: 0044421, GO: 0005615, GO: 0005576, GO: 1903561, GO: 0043230, and GO: 0070062) of the cellular component all turn out to describe the extracellular matrix and their interactions' contribution on the physical function enrichment. According to recent publications, both the internal secretion and external secretion of the pancreatic tissues can identify specific extracellular structures,[38,39] making such six extracellular items to enrich abundant optimal genes. Besides, a specific GO term in the biological process, tube development (GO: 0035295), was also enriched by optimal genes. Considering the complicated duct structure and cellular component of pancreatic tissues,[40,41] it is quite reasonable to enrich multiple duct cell-associated genes in such GO term, reflecting the specific biological functions of such cell subgroup.

For KEGG pathways, a specific pathway named pancreatic secretion (hsa04972) has been identified with an FDR of 3.14×10^{-15}. The secretion of pancreatic tissues, either internal or external, may be quite significant and involve certain subtypes of cells. Therefore, it is no doubt that the optimal genes may enrich in such biological pathway.

As mentioned above, the optimal genes enriched several GO terms and KEGG pathways that may reflect a specific a biological function for a specific cell subtype, confirming that these genes were quite essential for depicting the differences between pancreatic cells in different groups.

3.1.2. Analysis of Most Important Optimal Features. In this study, we obtained 655 optimal features. However, it is impossible to analyze the biological importance of each feature. By carefully checking the predicted results of SVM, we found that the MCC reached 0.874 when the top 10 features were used. Thus, we only analyzed the top 10 feature genes, which are listed in Table 3. The heat map of these 10 genes is shown

Table 3. Top Genes in the Feature List Yielded by MCFS Method

rank	gene symbol	RI
1	*SST*	7.20×10^{-1}
2	*INS*	7.04×10^{-1}
3	*PCSK1N*	6.77×10^{-1}
4	*CPA2*	6.63×10^{-1}
5	*REG1A*	6.27×10^{-1}
6	*GCG*	6.21×10^{-1}
7	*DCN*	6.06×10^{-1}
8	*TTR*	6.03×10^{-1}
9	*SCG5*	5.82×10^{-1}
10	*TRY6*	5.71×10^{-1}

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

in Figure 7, from which we can see that *PCSK1N*, *GCG*, *TTR*, and *SCG5* were expressed in alpha cells, *TRY6*, *CPA2*, and *REG1A* were expressed in acinar cells, *INS* was expressed in beta cells, *SST* was expressed in delta cells, and *DCN* was expressed in mesenchymal cells.

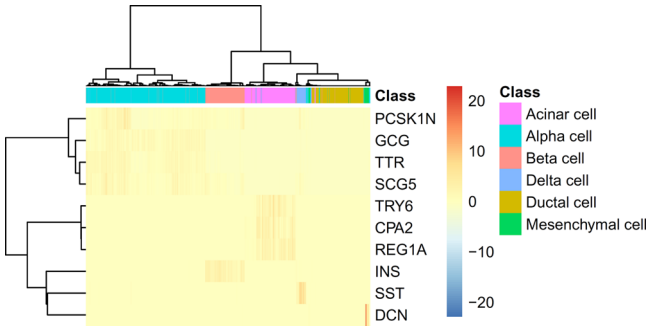


Figure 7. Heat map of the top 10 genes in six subgroups.

The top gene that may have a differential expression pattern in six cell subgroups was *SST*. Produced by encoding somatostatin, a functional inhibitory hormone, *SST* is a significant regulator interacting with multiple hormones of the gastrointestinal tract, including insulin[42] and glucagon.[43] For its differential expression patterns in six cell subgroups, it is highly expressed in the brain, gut, and delta cells in the pancreas,[42] implying its potential as a biomarker to distinguish pancreatic delta cells from the other candidate cells.

The next gene was *INS*, which is produced by encoding insulin. Insulin has recently been confirmed to be secreted by beta cells in the pancreas.[44−46] Therefore, as a transcriptomic marker, it may have a higher expression level in beta cells compared with the other five cell types.

The third gene, *PCSK1N*, encodes an inhibitor of prohormone convertase 1 and participates in the regulation of the proteolytic cleavage of neuroendocrine peptide precursors as the downstream of functional gene *PAX6*.[47] Considering that *PAX6* is the core regulator that contributes to the differentiation of pancreatic alpha, beta, and delta cells, *PAX6* may have a specific expression pattern in such three subgroups of cells.[48] Therefore, *PCSK1N*, as the downstream of *PAX6*, may also have its specific expression level under the control of *PAX6* in pancreatic alpha, beta, and delta cells, distinguishing such three subgroups of cells from cells in the other subgroups.[47]

Pancreatic specific carboxypeptidase A2 is encoded by the fourth gene, *CPA2*, which is a potential biomarker for cell clustering and recognition. As a member of the pancreatic carboxypeptidase family, such gene has a higher expression pattern in the exocrine regulatory cells (acinar and ductal cells) and mast cells,[49−51] distinguishing such cells from the other endocrine regulatory cells.

The fifth gene was *REG1A*, a type I subclass member of the Reg gene family, encoding an exocrine pancreas-secreting protein.[52] For its differential expression pattern in the six cell subgroups, considering that genes from the REG family contribute to the regeneration of pancreatic islet cells,[53−55] such gene may have a differential expression pattern in cells from endocrine regions (alpha, beta, and delta cells) compared with the other pancreatic cells.

The sixth gene, *GCG*, encodes glucagon, the antagonist of insulin.[56] Such hormone is also only secreted by a specific type of cells, alpha cells in the pancreas.[20,57−59] Therefore, similar to

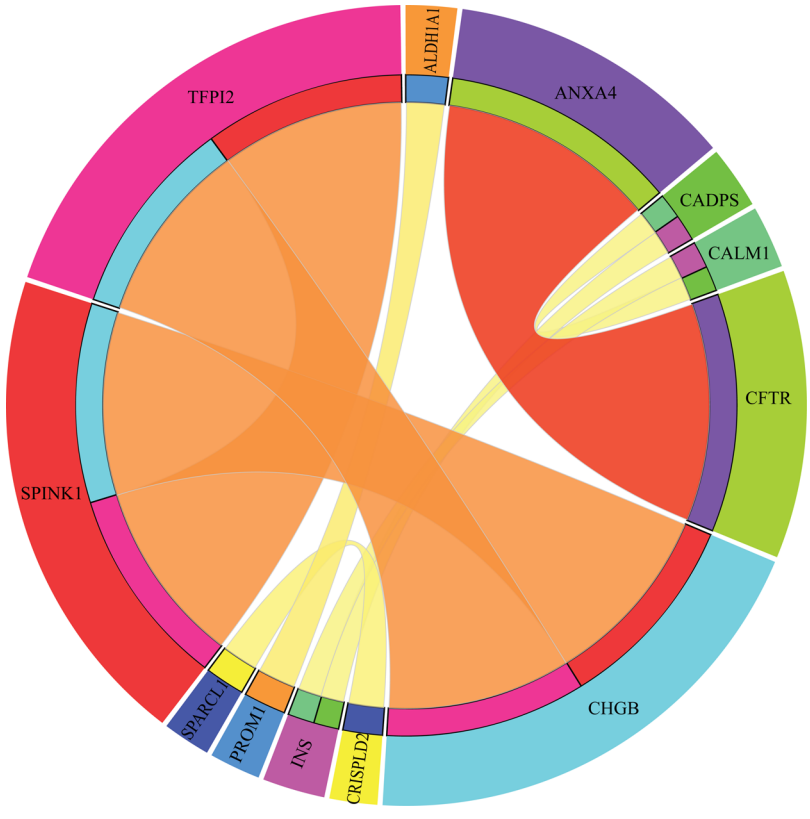


Figure 8. Rule networks for 12 classification rules generated by Ciruvis.[70]

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

when *INS* encodes insulin, such gene encoding glucagon can also tell the differences between the six cell subgroups at the transcriptomic level.

The seventh gene is *DCN*, encoding a small leucine-rich proteoglycan protein named decorin.[60,61] As for its cell subgroup-restricted expression patterns, the protein it encodes is a extracellular matrix regulator secreted by mesenchymal cells in the microenvironment of physical or pathological pancreas, implying that such gene may have differential expression pattern in mesenchymal cells.[60−62]

The eighth gene, named *TTR*, has also been confirmed to contribute to the regulation of glucose metabolism, a quite essential biological function of pancreatic endocrine regions.[63,64] This gene may directly be involved in an insulin-mediated glucose regulatory approach[64,65] and has quite a specific expression pattern in pancreatic alpha cells compared with the other cell subgroups.[66]

The ninth gene, *SCG5*, encodes a specific chaperone protein that prevents the aggregation of other secreted proteins.[67] For its differential expression pattern in the six cell subgroups, it contributes to prohormone processing and insulin exocytosis in the pancreas.[68] Therefore, it has a quite higher expression level in beta cells compared with other cell subgroups.

The last gene in the top 10 optimal genes, *TRY6*, also participates in the cleavage of functional peptides, similar to *PCSK1N*, and may directly regulate cell migration in the pancreas.[69] Furthermore, it contributes to the regulation of glucose metabolism by interfering with the biological function of beta cells in the pancreas,[69] indicating its specific expression pattern in this cell subgroup.

According to the above analysis, all top 10 genes in the optimal feature set have a differential expression pattern in the six cell subgroups. These genes may not only be potential biomarkers for pancreatic cell subtyping but may also contribute to the disclosure of the potential functional characteristics and distributions of pancreatic cellular components. Studies on the rest of the genes were left to the readers.

3.2. Analysis of Classification Rules. Aside from extracting important genes for distinguishing cells in the six subgroups, we also obtained 12 classification rules (listed in Table 2) for accurate recognition and clustering of pancreatic cells according to their specific expression patterns at the single-cell level. A rule network of these rules was plotted by Ciruvis,[70] as shown in Figure 8, from which it is easy to observe the internal interactions among the components of some rules and potential crosstalk between different rules. Genes *INS*, *CADPS*, and *CALM1* have a relative expression pattern and may interact with each other in physical conditions.[71,72] According to recent publications,[71,72] such speculation can be confirmed, forming a functional relatively "loop". Apart from that, similar "loops" constructed by some genes may also hide under the complicated quantitative interaction-based rules. Therefore, it can be implied that these quantitative rules may not only provide a novel way for accurate identification of complicated cell subtypes in the pancreas but also reflect the inner interaction relationship between some essential genes.[35−41] In addition, 16 genes were involved in the constructed rules. We plotted a heat map of them, as shown in Figure 9. It can be seen that the cell types with large sample sizes including acinar, alpha, beta, delta, and ductal cells were well clustered. Mesenchymal cells formed a small cluster and mixed with other scattered cells.

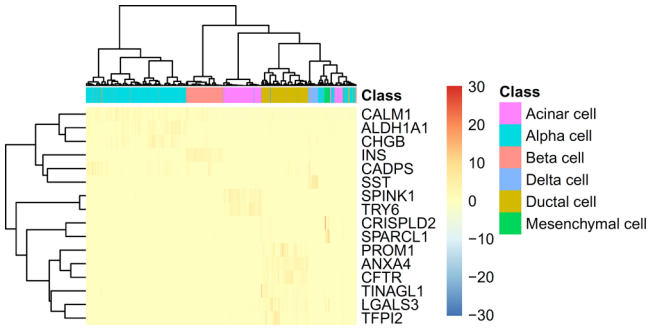


Figure 9. Heat map of 16 genes that were used to construct classification rules in six subgroups.

In the following text, we would analyze rules one by one. According to recent publications, either all rules can be precisely obeyed or the tendency can be predicted, validating the reliability of our results.

Only one rule (rule-1) contributes to the identification of mesenchymal cells, which used two functional genes: *SPARCL1* and *CRISPLD2*. On the basis of this rule, if both genes are expressed in the objective cell, then the candidate cell is a mesenchymal cell. *SPARCL1* participates in the maintenance of mesenchymal status.[73] Considering that only mesenchymal cells have a mesenchymal status among the six cell subgroups, it is quite reasonable to build such *SPARCL1*-related rule for cell subtyping. Apart from *SPARCL1*, *CRISPLD2* is a secondary parameter for the recognition of mesenchymal cells. This gene encodes a specific mesenchymal secretory protein,[74] indicating that among the six cell subgroups, only mesenchymal cells can have an optimal higher expression level on such gene, corresponding with this rule. Thus, the combined application of such two mesenchymal cell-specific markers may be quite effective and accurate for the identification of mesenchymal cells in the pancreas.

The second and third rules (rule-2 and rule-3) both contribute to the identification of pancreatic delta cells. It is worth noting that these two rules were all based on *SST*. As analyzed in Section 3.1, *SST* is highly expressed in delta cells of the pancreas.[42] As for the detailed threshold, according to a recent RNA sequencing publication involving the expression profiling of pancreatic cells provided by the GEO database,[75] the expression levels of *SST* in multiple cells, except in delta cells, are lower than 0.1 FPKM, partly validating the reasonability of these two rules.

The fourth and fifth rules (rule-4 and rule-5) contribute to the recognition of pancreatic beta cells. Both rules involve *INS* as the common gene. It encodes insulin, which is a quite significant marker for pancreatic beta cells, corresponding with both the threshold and alternative tendency.[44−46] Apart from *INS*, the lower expression level of *CALM1* and the higher expression level of *CADPS* contribute to the identification of pancreatic beta cells. *CALM1* is lowly expressed in the pancreas and only upregulated during the initiation and progression of pancreatic cancer.[71] Therefore, its low expression level (≤ -0.601) may correspond with the exact expression profiling of pancreatic beta cells. As for *CADPS* in rule-5, it is highly expressed in pancreatic beta cells, mediating the calcium-dependent secretion of insulin.[76,77] Considering the calcium-dependent insulin secretion mechanisms and core contribution of *CADPS* on calcium-dependent secretion, it is quite

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

reasonable to summarize that it may have a quite proper expression level in pancreatic beta cells.

The next four rules (rule-6, rule-7, rule-8, and rule-9) contribute to the identification of pancreatic ductal cells. Eight genes were used to construct these rules for quantitatively identifying pancreatic ductal cells at the single-cell level. In rule-6, the high expression levels of *CFTR* and *ANXA4* are deemed enough for cell subtype clustering. According to a recent publication,[78] the high expression level of *CFTR* contributes to the regulation of the ductal electrolyte and fluid transporters in the pancreas, indicating its specific expression pattern in ductal cells. Apart from *CFTR*, *ANXA4* encodes a calcium-dependent phospholipid binding protein, which is highly expressed in pancreatic ductal cells, corresponding with this rule.[23] In rule-7, three genes were involved. *TFPI2* encodes a specific inhibitor of the Kunitz-type serine proteinase and contributes to the malignant proliferation of only pancreatic ductal cells rather than the other pancreatic cells.[79] Different with *TFPI2*, in pancreatic ductal cancer, the expression levels of *SPINK1* and *CHGB* are both quite low,[80−82] corresponding with this rule. Therefore, the high expression level of *TFPI2* and low expression levels of *SPINK1* and *CHGB* may contribute to the identification of pancreatic ductal cells. The next rule (rule-8) was constructed using *TINAGL1*, which encodes a specific protein similar to tubulointerstitial nephritis antigen. This gene has been identified in the pancreas, especially in the pancreatic ductal tissue, where the pancreatic ductal adenocarcinoma is derived.[83] Therefore, its high expression level ($\geq$0.617) may also be a quantitative marker for pancreatic cell subtyping. For the last rule (rule-9) for pancreatic ductal cell identification, the high expression level of *PROM1* and low expression level of *ALDH1A1* have both been confirmed in normal pancreatic ductal cells according to recent studies[84−86] on gene expression profiling of pancreatic ductal cells under physical or pathological conditions, validating its reasonability.

Two rules (rule-10 and rule-11) were built for the identification of acinar cells. The high expression levels of *TRY6* and *LGALS3* contribute to the identification of acinar cells. As analyzed in Section 3.1, *TRY6* has a differential expression pattern in different cell subgroups.[69] Its high expression level indicates that our objectives may be functionally associated with exocrine regulatory cells (sugar metabolism regulation) but not with endocrine regulatory cells.[69] Therefore, such expression pattern may distinguish the remaining two subgroups of cells: acinar and alpha cells (endocrine regulatory cells). Similarly, the remaining rule restricted by *LGALS3* contributes to acinar cell identification. Galectin-3, which is encoded by *LGALS3*, directly contributes to the regulation and maintenance of normal cell biological processes in pancreatic exocrine acinar cells.[87,88] Finally, cells that do not fit with all aforementioned rules can be clustered into pancreatic alpha cells.

As analyzed above, the reasonability of all classification rules can be confirmed by recent publications. Combining the important genes analyzed in Section 3.1, the identified markers, either quantitative or qualitative, may not only contribute to the accurate recognition of complicated cell components in the pancreas at the single-cell transcriptome level but also provide a new tool for the disclosure of detailed biological functions of each cell subgroup and the molecular mechanisms in the pancreas under either physical or pathological conditions.

3.3. Comparison of Our Identified Pancreatic Markers at Single-Cell Level with Previous Studies. With the development of single-cell RNA sequencing technologies and related computational approaches,[27] various studies[33,89−92] have contributed to the identification of cell type-specific expression markers at single-cell resolution. In these studies, some biomarkers have already not only been identified but also been confirmed by the following experiments. Comparing the pancreatic cell biomarkers identified by such previous studies with those identified via our computational approach, several biomarkers identified in this study were also reported in these previous studies, partly validating the efficacy and accuracy of our approach.

Early in 2015, specific biomarkers such as *INS*, *REG1A*, and *SST* have already been identified as specific biomarkers for a certain subgroup of normal pancreatic cells, corresponding with our prediction.[33] However, such study only revealed some expression patterns of certain cell subtypes and did not try to distinguish such cell subgroups with quantitative rules. Following such study, another similar study[89] revealed the detailed cell components of pancreatic tissues and their subtype-specific biomarker genes. In this study, with the application of SORT-Seq (SOrting and Robot-assisted Transcriptome SEQuencing), researchers identified various cell type-specific biomarkers at the single-cell transcriptomic level. Genes such as *INS* and *SST* have been directly confirmed to be expressed in beta and delta cells, respectively, which also corresponds with our results. However, for each cell subtype, such study identified more than 20 biomarkers, which no doubt indicate huge redundancy. The third study[90] provided by Baron et al. also contributed to the identification of the pancreatic population structure at the level of single-cell transcriptome. For the first time, they not only focused on the pancreatic tissue cells such as alpha and beta cells but also paid attention to related immune cells and functional myeloid cells in the microenvironment. Compared with our results, in such study, they only used one biomarker for the clustering of each cell subtype, such as *INS* in beta cells. Such identified biomarkers, such as *INS*, *SST*, and *CPA1*, have all been predicted and screened out in our study.[90] Using more biomarkers for one cell subtype may further improve the efficacy and accuracy. Further in 2016, the study[91] presented by Segerstolpe et al. compared the single-cell expression profiling in physical and pathogenic conditions. In such study, specific biomarkers for alpha, beta, and acinar cells have also been identified. Comparing such results with our results, various biomarkers have also been shared, such as *GCG*, *INS*, and *SST*. Similar with Segerstolpe et al.'s study, in 2017, Lawlor et al.[92] further compared the specific biomarkers for pancreatic cell subgroups at the single-cell level under either physical or pathological (type 2 diabetes) conditions, sharing similar biomarkers for pancreatic alpha, beta, and gamma cells with our results.

As mentioned above, compared with previous studies, several biomarkers reported in our study were consistent with those in previous experiment-based studies. However, for the first time, we presented a quantitative analytic approach for pancreatic cell subtyping, filling the gap for previous qualitative analysis.

4. MATERIALS AND METHODS

4.1. Datasets. We downloaded the single-cell RNA-seq data of the human pancreas from the Gene Expression

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

Omnibus (GEO) database under accession number GSE81547.[26] There were 2282 cells from six known cell subgroups: (1) acinar, (2) alpha, (3) beta, (4) delta, (5) ductal, and (6) mesenchymal cells. The number of cells in each subgroup is listed in Table 4. For each cell, the expression levels of 23,459 genes were measured with RNA sequencing. The purpose of this study was to identify the genes that were differentially expressed among different cell types and obtain the pancreatic cell biomarkers.

Table 4. Number of Cells in the Six Subgroups

cell type	number of cells
acinar cell	411
alpha cell	998
beta cell	348
delta cell	83
ductal cell	389
mesenchymal cell	53
total	2282

4.2. Feature Selection Method. As mentioned in Section 4.1, each cell was represented by the expression profiling data. To analyze these data, we applied a two-step feature selection procedure to identify discriminate genes associated with the six subgroups of cells.

First, the MCFS method was applied to analyze the expression profiling data.[28] As mentioned in Section 4.1, there were 2282 cells, and each cell was represented by 23,459 features, meaning that we had to deal with a dataset with high dimensionality and small sample size. As elaborated by Draminski et al.[28] and confirmed in some studies,[93−95] the MCFS method is good at analyzing this type of dataset and capturing essential information. Thus, it is quite proper to deal with expression profiling data. To date, it has been applied to tackle several biological problems.[93−101] MCFS is a supervised feature selection method, which grows multiple decision trees from bootstrap samples and feature subsets with m features that are randomly selected from N original features ($m \ll N$). Its brief description is as follows, and the detailed introduction can be seen in Draminski et al.'s study.[28]

1. t feature subsets with m features are randomly generated from original N features.
2. For each feature subset, p decision trees are grown from p bootstrap training sets, whose samples are represented using the features in this feature subset.
3. Step 2 is repeated for the remaining $t - 1$ feature subsets.
4. In total, we obtain $p \times t$ decision trees.

To yield the importance of each input feature, relative importance (RI) is defined according to how a feature is involved in each constructed decision tree classifier. The more frequent the feature is involved, the higher the RI score of this feature is.

On the basis of the derived RI scores for all features, a feature list with decreasing RI scores can be obtained, which is formulated as

$$F = [f_1, f_2, ..., f_N] \quad (1)$$

where N is the number of features used in this study. Here, the program of the MCFS method was retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html (dmLab, version 2.1.1). For convenience, we used its default parameters. In detail, two regular factors, u and v, were all set to 1.0.

Second, the IFS[29] method was further used to select discriminate features with the help of a supervised classifier. The original IFS method always tests all possible feature subsets, thereby extracting the most important features. However, 23,459 features were considered in this study, and the possible feature sets are too many, causing us to require a large amount of time to complete the test tasks due to our limited computational power. Here, we applied a two-stage IFS method. In the first stage, a series of feature subsets with step 10 were picked up, and a classifier was employed to test them one by one, inducing a possible number range where the most important feature set may be located. In the second stage, all feature subsets with step 1 in this range were further tested by the selected classifier, and the feature subset that can yield the best performance can be obtained. The obtained feature subset was called the optimal feature set, and features in this set were termed optimal features. In addition, the classifier based on the optimal features was named the optimal classifier.

4.3. Support Vector Machine. SVM is a widely used supervised classifier based on the statistical learning theory; it finds a hyperplane with a maximum margin between samples in two different classes. SVMs are applied in many biological problems for both binary and multiclass classification.[93,102−107] For nonlinear data, the kernel trick is used to map the data in a nonlinear, low-dimensional space to a linear, high-dimensional space. For multiclass classification, the "One Versus the Rest" strategy is adopted that trains C binary SVM classifiers for C classes, where each classifier is trained using the samples of that class as positives and samples from other classes as negatives.

To quickly implement the SVM, we adopted a tool "SMO" in Weka,[108] a suite of software collecting a large number of widely used machine learning algorithms and data process tools. This tool implements the SVM optimized by sequential minimum optimization (SMO).[109] For convenience, it was executed using its default parameters. In detail, the kernel was set to be a polynomial function, and tolerance parameter was set to 0.001.

4.4. Rule Learning. As mentioned in Section 4.2, the MCFS method can evaluate the importance of each feature using the RI score. Furthermore, it can select some features with highest RI scores, that is, some top features in list F, as informative features. The critical value for picking up informative features is determined by a permutation test on class labels and a following one-sided Student's t-test.[110] Features with RI scores larger than such critical value are deemed informative features, which are considered to be essential for the classification problem.

On the basis of obtained informative features, some classification rules can be extracted, which can give a clear picture for classifying a given pancreatic cell into six subtypes, thereby improving the comprehension on differences of cells in different subtypes. To do that, the Johnson Reducer algorithm[31] was first applied on informative features to produce a reduced feature set that contained most important informative features and can produce similar classification ability compared with using all informative features. Then, a rule learning algorithm, RIPPER,[32] was adopted to generate classification rules using features in the reduced set. Rules yielded by the RIPPER algorithm always contain two parts: (1) conditions and (2) outcome. For instance, IF Gene1 $\leq$ 0.125 AND Gene2 $\geq$ 3.102 THEN Acinar cell. The "Gene1 $\leq$ 0.125

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

AND Gene2 ≥ 3.102" was the condition, and "Acinar cell" was the outcome. Fortunately, the program of the MCFS method also integrated the above procedures; that is, it can further produce the classification rules besides evaluating the importance of features. The obtained classification rules would be directly used in this study to uncover the gene expression differences on pancreatic cells in the six subtypes.

4.5. Performance Measurement. In this study, the 10-fold cross-validation strategy[93,96,98,104,105,111−116] was used to evaluate the performance of trained multiclass classifiers. In this strategy, all samples are randomly and equally divided into 10 sets. The procedures contain 10 rounds. In each round, samples in one set are deemed test samples, and the others are used to train the classifier. After 10 rounds, each sample is tested exactly once. By collecting and counting the predicted class of each sample, several measurements are obtained to quantify the performance of the classifier. This evaluation strategy is more popular than the jackknife test[117,118] because it can save a large amount of time and yield similar results.

As mentioned in Section 4.1, six subgroups of cells were considered. For the predicted results yielded by a classifier, we can compute the prediction accuracy for each subgroup, which is defined as the proportion of correctly predicted cells and all cells in one subgroup. In addition, the overall accuracy (ACC) can be calculated by

$$\mathrm{ACC} = \frac{\sum_{i=1}^{6} C_i}{\sum_{i=1}^{6} N_i} \tag{2}$$

where C_i represents the number of correctly predicted cells, and N_i denotes the total number of cells in the ith subgroup.

In addition, on the basis of the subgroup sizes listed in Table 1, the dataset was imbalanced. The largest group had 998 cells, whereas the smallest group had 53 cells. In this case, the ACC cannot indicate the performance of a classifier on the whole. Thus, we employed the Matthew's correlation coefficient (MCC),[119] which is deemed a balanced measurement. However, the original MCC applies for binary classification problems. Thus, for each subgroup, we can compute its MCC as follows

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{3}$$

where TP represents the number of corrected predicted cells in this subgroup, FN is the number of incorrectly predicted cells in this subgroup, FP is the number of cells that are in other subgroups and predicted to be in such subgroup, and TN is the number of cells that are in other subgroups and not predicted to be in such subgroup.

However, the above-mentioned MCCs cannot evaluate the overall performance of a classifier. Thus, we further employed the multiclass version of MCC, which was proposed by Gorodkin.[120] The following text gives its brief description.

Suppose we have N samples, denoted as $s_1, s_2,...,s_N$, and C classes, represented by 1,2,...,C. Let $X = (x_{ij})_{N \times C}$ be a matrix representing the predicted classes of all samples, where x_{ij} is a binary value (x_{ij} is 1 if s_i is accurately predicted as class j; otherwise, x_{ij} is 0). Similarly, the matrix $Y = (y_{ij})_{n \times C}$ is a binary matrix defining the true classes of samples. Then, the MCC in multiclass can be calculated using the following formula

$$\mathrm{MCC} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)\mathrm{cov}(Y, Y)}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}\sum_{j=1}^{C}(y_{ij} - \bar{y}_j)^2}} \tag{4}$$

where $\bar{x}_j$ and $\bar{y}_j$ are the mean values of x_j and y_j, respectively.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.8b02171.

Feature list yielded by the MCFS method, performance of SVM under feature sets containing a multiple of 10 features, performance of SVM of feature sets containing 560−660 features, and performance of NNA under feature sets containing different numbers of features (PDF)

Enrichment analysis results on the top 655 genes (XLSX)

Independent test dataset to test the utility of classification rules (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: tohuangtao@126.com. Tel.: 0086-21-54923269 (T.H.).

*E-mail: cai_yud@126.com. Tel.: 0086-21-66136132. Fax: 0086-21-66136109 (Y.-D.C.).

ORCID

Lei Chen: 0000-0003-3068-1583

Tao Huang: 0000-0003-1975-9693

Yu-Dong Cai: 0000-0001-5664-7979

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR1412500), National Key R&D Program of China (2018YFC0910403), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the Key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), and Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

■ REFERENCES

(1) Ziv, O.; Glaser, B.; Dor, Y. The plastic pancreas. *Dev. Cell* **2013**, *26*, 3−7.

(2) Benitez, C. M.; Goodyer, W. R.; Kim, S. K. Deconstructing pancreas developmental biology. *Cold Spring Harbor Perspect. Biol.* **2012**, *4*, a012401.

(3) Cao, Z.; Wang, X. The endocrine role between beta cells and intra-islet endothelial cells. *Endocr. J.* **2014**, *61*, 647−654.

(4) Bharat, A.; Saini, D.; Benshoff, N.; Goodman, J.; Desai, N. M.; Chapman, W. C.; Mohanakumar, T. Role of intra-islet endothelial cells in islet allo-immunity. *Transplantation* **2007**, *84*, 1316−1323.

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

(5) Pezzilli, R. Exocrine pancreas involvement in celiac disease: a review. *Recent Pat. Inflammation Allergy Drug Discovery* **2014**, *8*, 167−172.
(6) Osundiji, M. A.; Evans, M. L. Brain control of insulin and glucagon secretion. *Endocrinol. Metab. Clin. North Am.* **2013**, *42*, 1−14.
(7) Pandiri, A. R. Overview of exocrine pancreatic pathobiology. *Toxicol. Pathol.* **2014**, *42*, 207−216.
(8) Collombat, P.; Hecksher-Sorensen, J.; Krull, J.; Berger, J.; Riedel, D.; Herrera, P. L.; Serup, P.; Mansouri, A. Embryonic endocrine pancreas and mature beta cells acquire alpha and PP cell phenotypes upon Arx misexpression. *J. Clin. Invest.* **2007**, *117*, 961−970.
(9) Habener, J. F.; Stanojevic, V. Alpha cells come of age. *Trends Endocrinol. Metab.* **2013**, *24*, 153−163.
(10) Rodriguez-Diaz, R.; Caicedo, A. Neural control of the endocrine pancreas. *Best Pract. Res., Clin. Endocrinol. Metab.* **2014**, *28*, 745−756.
(11) Mizgier, M. L.; Casas, M.; Contreras-Ferrat, A.; Llanos, P.; Galgani, J. E. Potential role of skeletal muscle glucose metabolism on the regulation of insulin secretion. *Obes. Rev.* **2014**, *15*, 587−597.
(12) Said, H. M. Recent advances in transport of water-soluble vitamins in organs of the digestive system: a focus on the colon and the pancreas. *Am. J. Physiol.* **2013**, *305*, G601−610.
(13) Feig, C.; Gopinathan, A.; Neesse, A.; Chan, D. S.; Cook, N.; Tuveson, D. A. The pancreas cancer microenvironment. *Clin. Cancer Res.* **2012**, *18*, 4266−4276.
(14) Bijlsma, M. F.; van Laarhoven, H. W. The conflicting roles of tumor stroma in pancreatic cancer and their contribution to the failure of clinical trials: a systematic review and critical appraisal. *Cancer Metastasis Rev.* **2015**, *34*, 97−114.
(15) Saliba, A. E.; Westermann, A. J.; Gorski, S. A.; Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.* **2014**, *42*, 8845−8860.
(16) Trombetta, J. J.; Gennert, D.; Lu, D.; Satija, R.; Shalek, A. K.; Regev, A. Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr. Protoc. Mol. Biol.* **2014**, *107*, 4.22.1−4.22.17.
(17) Zhao, X.; Gao, S.; Wu, Z.; Kajigaya, S.; Feng, X.; Liu, Q.; Townsley, D. M.; Cooper, J.; Chen, J.; Keyvanfar, K.; Fernandez Ibanez, M. D. P.; Wang, X.; Young, N. S. Single-cell RNA-seq reveals a distinct transcriptome signature of aneuploid hematopoietic cells. *Blood* **2017**, *130*, 2762−2773.
(18) Cho, D. S.; Doles, J. D. Single cell transcriptome analysis of muscle satellite cells reveals widespread transcriptional heterogeneity. *Gene* **2017**, *636*, 54−63.
(19) Cleveland, M. H.; Sawyer, J. M.; Afelik, S.; Jensen, J.; Leach, S. D. Exocrine ontogenies: on the development of pancreatic acinar, ductal and centroacinar cells. *Semin. Cell Dev. Biol.* **2012**, *23*, 711−719.
(20) Stanojevic, V.; Habener, J. F. Evolving function and potential of pancreatic alpha cells. *Best Pract. Res., Clin. Endocrinol. Metab.* **2015**, *29*, 859−871.
(21) Fu, A.; Eberhard, C. E.; Screaton, R. A. Role of AMPK in pancreatic beta cell function. *Mol. Cell. Endocrinol.* **2013**, *366*, 127−134.
(22) Brereton, M. F.; Vergari, E.; Zhang, Q.; Clark, A. Alpha-, Delta- and PP-cells: Are They the Architectural Cornerstones of Islet Structure and Co-ordination? *J. Histochem. Cytochem.* **2015**, *63*, 575−591.
(23) Maleth, J.; Hegyi, P. Calcium signaling in pancreatic ductal epithelial cells: an old friend and a nasty enemy. *Cell Calcium* **2014**, *55*, 337−345.
(24) Huang, M.; Xin, W. Matrine inhibiting pancreatic cells epithelial-mesenchymal transition and invasion through ROS/NF-kappaB/MMPs pathway. *Life Sci.* **2018**, *192*, 55−61.
(25) Wu, Y. S.; Chung, I.; Wong, W. F.; Masamune, A.; Sim, M. S.; Looi, C. Y. Paracrine IL-6 signaling mediates the effects of pancreatic stellate cells on epithelial-mesenchymal transition via Stat3/Nrf2 pathway in pancreatic cancer cells. *Biochim. Biophys. Acta* **2017**, *1861*, 296−306.
(26) Enge, M.; Arda, H. E.; Mignardi, M.; Beausang, J.; Bottino, R.; Kim, S. K.; Quake, S. R. Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* **2017**, *171*, 321−330 e14 .
(27) Soneson, C.; Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **2018**, *15*, 255−261.
(28) Draminski, M.; Rada-Iglesias, A.; Enroth, S.; Wadelius, C.; Koronacki, J.; Komorowski, J. Monte Carlo feature selection for supervised classification. *Bioinformatics* **2008**, *24*, 110−117.
(29) Liu, H. A.; Setiono, R. Incremental feature selection. *Appl. Intell.* **1998**, *9*, 217−230.
(30) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273−297.
(31) Johnson, D. S. Approximation algorithms for combinatorial problems. *J. Comp. Syst. Sci.* **1974**, *9*, 256−278.
(32) Cohen, W. W. Fast effective rule induction. In *The Twelfth International Conference On Machine Learning*, 1995; pp 115-123.
(33) Li, J.; Klughammer, J.; Farlik, M.; Penz, T.; Spittler, A.; Barbieux, C.; Berishvili, E.; Bock, C.; Kubicek, S. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO Rep* **2016**, *17*, 178−187.
(34) Li, J.; Casteels, T.; Frogne, T.; Ingvorsen, C.; Honore, C.; Courtney, M.; Huber, K. V. M.; Schmitner, N.; Kimmel, R. A.; Romanov, R. A.; Sturtzel, C.; Lardeau, C.-H.; Klughammer, J.; Farlik, M.; Sdelci, S.; Vieira, A.; Avolio, F.; Briand, F.; Baburin, I.; Majek, P.; Pauler, F. M.; Penz, T.; Stukalov, A.; Gridling, M.; Parapatics, K.; Barbieux, C.; Berishvili, E.; Spittler, A.; Colinge, J.; Bennett, K. L.; Hering, S.; Sulpice, T.; Bock, C.; Distel, M.; Harkany, T.; Meyer, D.; Superti-Furga, G.; Collombat, P.; Hecksher-Sorensen, J.; Kubicek, S. Artemisinins Target GABAA Receptor Signaling and Impair alpha Cell Identity. *Cell* **2017**, *168*, 86−100.
(35) Senoo, H.; Mezaki, Y.; Fujiwara, M. The stellate cell system (vitamin A-storing cell system). *Anat. Sci. Int.* **2017**, *92*, 387−455.
(36) Farney, A. C.; Rogers, J.; Stratta, R. J. Pancreas graft thrombosis: causes, prevention, diagnosis, and intervention. *Curr. Opin. Organ Transplant.* **2012**, *17*, 87−92.
(37) Lyman, G. H.; Bohlke, K.; Khorana, A. A.; Kuderer, N. M.; Lee, A. Y.; Arcelus, J. I.; Balaban, E. P.; Clarke, J. M.; Flowers, C. R.; Francis, C. W.; Gates, L. E.; Kakkar, A. K.; Key, N. S.; Levine, M. N.; Liebman, H. A.; Tempero, M. A.; Wong, S. L.; Somerfield, M. R.; Falanga, A. Venous thromboembolism prophylaxis and treatment in patients with cancer: american society of clinical oncology clinical practice guideline update 2014. *J. Clin. Oncol.* **2015**, *33*, 654−656.
(38) Bynigeri, R. R.; Jakkampudi, A.; Jangala, R.; Subramanyam, C.; Sasikala, M.; Rao, G. V.; Reddy, D. N.; Talukdar, R. Pancreatic stellate cell: Pandora's box for pancreatic disease biology. *World J. Gastroenterol.* **2017**, *23*, 382−405.
(39) Leppkes, M.; Maueroder, C.; Hirth, S.; Nowecki, S.; Günther, C.; Billmeier, U.; Paulus, S.; Biermann, M.; Munoz, L. E.; Hoffmann, M.; Wildner, D.; Croxford, A. L.; Waisman, A.; Mowen, K.; Jenne, D. E.; Krenn, V.; Mayerle, J.; Lerch, M. M.; Schett, G.; Wirtz, S.; Neurath, M. F.; Herrmann, M.; Becker, C. Externalized decondensed neutrophil chromatin occludes pancreatic ducts and drives pancreatitis. *Nat. Commun.* **2016**, *7*, 10973.
(40) Zare, M.; Rastegar, S.; Ebrahimi, E.; Roohipoor, A.; Shirali, S. Role of pancreatic duct cell in beta cell neogenesis: A mini review study. *Diabetes Metab. Syndr.: Clin Res. Rev.* **2017**, *11*, S1−S4.
(41) Hayashi, M.; Novak, I. Molecular basis of potassium channels in pancreatic duct epithelial cells. *Channels* **2013**, *7*, 432−441.
(42) Braun, M. The somatostatin receptor in human pancreatic beta-cells. *Vitam. Horm.* **2014**, *95*, 165−193.
(43) Salvatori, A. S.; Elrick, M. M.; Samson, W. K.; Corbett, J. A.; Yosten, G. L. C. Neuronostatin inhibits glucose-stimulated insulin secretion via direct action on the pancreatic alpha-cell. *Am. J. Physiol. Endocrinol. Metab.* **2014**, *306*, E1257−1263.

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

(44) Hardy, A. B.; Prentice, K. J.; Froese, S.; Liu, Y.; Andrews, G. K.; Wheeler, M. B. Zip4 mediated zinc influx stimulates insulin secretion in pancreatic beta cells. *PLoS One* **2015**, *10*, No. e0119136.
(45) Fu, Z.; Gilbert, E. R.; Liu, D. Regulation of insulin synthesis and secretion and pancreatic Beta-cell dysfunction in diabetes. *Curr. Diabetes Rev.* **2013**, *9*, 25−53.
(46) Rutter, G. A.; Pullen, T. J.; Hodson, D. J.; Martinez-Sanchez, A. Pancreatic beta-cell identity, glucose sensing and the control of insulin secretion. *Biochem. J.* **2015**, *466*, 203−218.
(47) Liu, T.; Zhao, Y.; Tang, N.; Feng, R.; Yang, X.; Lu, N.; Wen, J.; Li, L. Pax6 directly down-regulates Pcsk1n expression thereby regulating PC1/3 dependent proinsulin processing. *PLoS One* **2012**, *7*, No. e46934.
(48) Xie, Q.; Yang, Y.; Huang, J.; Ninkovic, J.; Walcher, T.; Wolf, L.; Vitenzon, A.; Zheng, D.; Gotz, M.; Beebe, D. C.; Zavadil, J.; Cvekl, A. Pax6 interactions with chromatin and identification of its novel direct target genes in lens and forebrain. *PLoS One* **2013**, *8*, No. e54507.
(49) Nakano, E.; Geisz, A.; Masamune, A.; Niihori, T.; Hamada, S.; Kume, K.; Kakuta, Y.; Aoki, Y.; Matsubara, Y.; Ebert, K.; Ludwig, M.; Braun, M.; Groneberg, D. A.; Shimosegawa, T.; Sahin-Tóth, M.; Witt, H. Variants in pancreatic carboxypeptidase genes CPA2 and CPB1 are not associated with chronic pancreatitis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **2015**, *309*, G688−G694.
(50) Shirasawa, S.; Yoshie, S.; Yue, F.; Ichikawa, H.; Yokoyama, T.; Nagai, M.; Tomotsune, D.; Hirayama, M.; Sasaki, K. Pancreatic exocrine enzyme-producing cell differentiation via embryoid bodies from human embryonic stem cells. *Biochem. Biophys. Res. Commun.* **2011**, *410*, 608−613.
(51) Ventura, S.; Gomis-Rüth, F. X.; Puigserver, A.; Avilés, F. X.; Vendrell, J. Pancreatic procarboxypeptidases: oligomeric structures and activation processes revisited. *Biol. Chem.* **1997**, *378*, 161−165.
(52) Yuan, R.-H.; Jeng, Y.-M.; Chen, H.-L.; Hsieh, F.-J.; Yang, C.-Y.; Lee, P.-H.; Hsu, H.-C. Opposite roles of human pancreatitis-associated protein and REG1A expression in hepatocellular carcinoma: association of pancreatitis-associated protein expression with low-stage hepatocellular carcinoma, beta-catenin mutation, and favorable prognosis. *Clin. Cancer Res.* **2005**, *11*, 2568−2575.
(53) Takasawa, S. Regenerating gene (REG) product and its potential clinical usage. *Expert Opin. Ther. Targets* **2016**, *20*, 541−550.
(54) Ota, H.; Itaya-Hironaka, A.; Yamauchi, A.; Sakuramoto-Tsuchida, S.; Miyaoka, T.; Fujimura, T.; Tsujinaka, H.; Yoshimoto, K.; Nakagawara, K.; Tamaki, S.; Takasawa, S.; Kimura, H. Pancreatic beta cell proliferation by intermittent hypoxia via up-regulation of Reg family genes and HGF gene. *Life Sci.* **2013**, *93*, 664−672.
(55) Kapur, R.; Højfeldt, T. W.; Højfeldt, T. W.; Ronn, S. G.; Karlsen, A. E.; Heller, R. S. Short-term effects of INGAP and Reg family peptides on the appearance of small beta-cells clusters in non-diabetic mice. *Islets* **2012**, *4*, 40−48.
(56) Wu, S.; Xiang, K.; Bell, G. I. Dinucleotide repeat polymorphism in the human glucagon gene (GCG). *Nucleic Acids Res.* **1991**, *19*, 1163.
(57) Mohan, R.; Mao, Y.; Zhang, S.; Zhang, Y.-W.; Xu, C.-R.; Gradwohl, G.; Tang, X. Differentially Expressed MicroRNA-483 Confers Distinct Functions in Pancreatic beta- and alpha-Cells. *J. Biol. Chem.* **2015**, *290*, 19955−19966.
(58) Heddad Masson, M.; Poisson, C.; Guérardel, A.; Mamin, A.; Philippe, J.; Gosmain, Y. Foxa1 and Foxa2 regulate alpha-cell differentiation, glucagon biosynthesis, and secretion. *Endocrinology* **2014**, *155*, 3781−3792.
(59) Gosmain, Y.; Cheyssac, C.; Heddad Masson, M.; Dibner, C.; Philippe, J. Glucagon gene expression in the endocrine pancreas: the role of the transcription factor Pax6 in alpha-cell differentiation, glucagon biosynthesis and secretion. *Diabetes, Obes. Metab.* **2011**, *13*, 31−38.
(60) Köninger, J.; Giese, N. A.; di Mola, F. F.; Berberat, P.; Giese, T.; Esposito, I.; Bachem, M. G.; Büchler, M. W.; Friess, H. Overexpressed decorin in pancreatic cancer: potential tumor growth inhibition and attenuation of chemotherapeutic action. *Clin. Cancer Res.* **2004**, *10*, 4776−4783.
(61) Köninger, J.; Giese, N. A.; Bartel, M.; di Mola, F. F.; Berberat, P. O.; di Sebastiano, P.; Giese, T.; Büchler, M. W.; Friess, H. The ECM proteoglycan decorin links desmoplasia and inflammation in chronic pancreatitis. *J. Clin. Pathol.* **2006**, *59*, 21−27.
(62) Kadirvel, R.; Ding, Y.-H.; Dai, D.; Lewis, D. A.; Kallmes, D. F. Differential expression of genes in elastase-induced saccular aneurysms with high and low aspect ratios. *Neurosurgery* **2010**, *66*, 578−584.
(63) Fruscalzo, A.; Londero, A. P.; Driul, L.; Henze, A.; Tonutti, L.; Ceraudo, M.; Zanotti, G.; Berni, R.; Schweigert, F. J.; Raila, J. First trimester concentrations of the TTR-RBP4-retinol complex components as early markers of insulin-treated gestational diabetes mellitus. *Clin. Chem. Lab. Med.* **2015**, *53*, 1643−1651.
(64) Dekki, N.; Refai, E.; Holmberg, R.; Köhler, M.; Jornvall, H.; Berggren, P.-O.; Juntti-Berggren, L. Transthyretin binds to glucose-regulated proteins and is subjected to endocytosis by the pancreatic beta-cell. *Cell. Mol. Life Sci.* **2012**, *69*, 1733−1743.
(65) Zemany, L.; Bhanot, S.; Peroni, O. D.; Murray, S. F.; Moraes-Vieira, P. M.; Castoldi, A.; Manchem, P.; Guo, S.; Monia, B. P.; Kahn, B. B. Transthyretin Antisense Oligonucleotides Lower Circulating RBP4 Levels and Improve Insulin Sensitivity in Obese Mice. *Diabetes* **2015**, *64*, 1603−1614.
(66) Dorrell, C.; Grompe, M. T.; Pan, F. C.; Zhong, Y.; Canaday, P. S.; Shultz, L. D.; Greiner, D. L.; Wright, C. V.; Streeter, P. R.; Grompe, M. Isolation of mouse pancreatic alpha, beta, duct and acinar populations with cell surface markers. *Mol. Cell. Endocrinol.* **2011**, *339*, 144−150.
(67) Martin, T. M.; Plautz, S. A.; Pannier, A. K. Network analysis of endogenous gene expression profiles after polyethyleneimine-mediated DNA delivery. *J. Gene Med.* **2013**, *15*, 142−154.
(68) Pepaj, M.; Bredahl, M. K.; Gjerlaugsen, N.; Thorsby, P. M. Proteomic analysis of the INS-1E secretome identify novel vitamin D-regulated proteins. *Diabetes/Metab. Res. Rev.* **2016**, *32*, 514−521.
(69) González, N.; Martín-Duce, A.; Martínez-Arrieta, F.; Moreno-Villegas, Z.; Portal-Núñez, S.; Sanz, R.; Egido, J. Effect of bombesin receptor subtype-3 and its synthetic agonist on signaling, glucose transport and metabolism in myocytes from patients with obesity and type 2 diabetes. *Int. J. Mol. Med.* **2015**, *35*, 925−931.
(70) Bornelöv, S.; Marillet, S.; Komorowski, J. Ciruvis: a web-based tool for rule networks and interaction detection using rule-based classifiers. *BMC Bioinf.* **2014**, *15*, 139.
(71) Song, D.; Chaerkady, R.; Tan, A. C.; García-García, E.; Nalli, A.; Suárez-Gauthier, A.; López-Ríos, F.; Zhang, X. F.; Solomon, A.; Tong, J.; Read, M.; Fritz, C.; Jimeno, A.; Pandey, A.; Hidalgo, M. Antitumor activity and molecular effects of the novel heat shock protein 90 inhibitor, IPI-504, in pancreatic cancer. *Mol. Cancer Ther.* **2008**, *7*, 3275−3284.
(72) Speidel, D.; Salehi, A.; Obermueller, S.; Lundquist, I.; Brose, N.; Renström, E.; Rorsman, P. CAPS1 and CAPS2 regulate stability and recruitment of insulin granules in mouse pancreatic beta cells. *Cell Metab.* **2008**, *7*, 57−67.
(73) Hu, H.; Zhang, H.; Ge, W.; Liu, X.; Loera, S.; Chu, P.; Chen, H.; Peng, J.; Zhou, L.; Yu, S.; Yuan, Y.; Zhang, S.; Lai, L.; Yen, Y.; Zheng, S. Secreted protein acidic and rich in cysteines-like 1 suppresses aggressiveness and predicts better survival in colorectal cancers. *Clin. Cancer Res.* **2012**, *18*, 5438−5448.
(74) Zhang, H.; Kho, A. T.; Wu, Q.; Halayko, A. J.; Limbert Rempel, K.; Chase, R. P.; Sweezey, N. B.; Weiss, S. T.; Kaplan, F. CRISPLD2 (LGL1) inhibits proinflammatory mediators in human fetal, adult, and COPD lung fibroblasts and epithelial cells. *Physiol. Rep.* **2016**, *4*, No. e12942.
(75) Barrett, T.; Wilhite, S. E.; Ledoux, P.; Evangelista, C.; Kim, I. F.; Tomashevsky, M.; Marshall, K. A.; Phillippy, K. H.; Sherman, P. M.; Holko, M.; Yefanov, A.; Lee, H.; Zhang, N.; Robertson, C. L.; Serova, N.; Davis, S.; Soboleva, A. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* **2013**, *41*, D991−D995.
(76) Yaluri, N.; Modi, S.; López Rodriguez, M.; Stančáková, A.; Kuusisto, J.; Kokkola, T.; Laakso, M. Simvastatin Impairs Insulin

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

Secretion by Multiple Mechanisms in MIN6 Cells. *PLoS One* **2015**, *10*, No. e0142902.
(77) Skrzypski, M.; Kakkassery, M.; Mergler, S.; Grötzinger, C.; Khajavi, N.; Sassek, M.; Szczepankiewicz, D.; Wiedenmann, B.; Nowak, K. W.; Strowski, M. Z. Activation of TRPV4 channel in pancreatic INS-1E beta cells enhances glucose-stimulated insulin secretion via calcium-dependent mechanisms. *FEBS Lett.* **2013**, *587*, 3281−3287.
(78) Pallagi, P.; Hegyi, P.; Rakonczay, Z., Jr. The Physiology and Pathophysiology of Pancreatic Ductal Secretion: The Background for Clinicians. *Pancreas* **2015**, *44*, 1211−1233.
(79) Sato, N.; Parker, A. R.; Fukushima, N.; Miyagi, Y.; Iacobuzio-Donahue, C. A.; Eshleman, J. R.; Goggins, M. Epigenetic inactivation of TFPI-2 as a common mechanism associated with growth and invasion of pancreatic ductal adenocarcinoma. *Oncogene* **2005**, *24*, 850−858.
(80) Schubert, S.; Traub, F.; Brakensiek, K.; von Kopylow, K.; Marohn, B.; Maelzer, M.; Gaedcke, J.; Kreipe, H.; Stuhrmann, M. CFTR, SPINK1, PRSS1, and CTRC mutations are not associated with pancreatic cancer in German patients. *Pancreas* **2014**, *43*, 1078−1082.
(81) Rani, S.; O'Driscoll, L. Reverse-transcriptase polymerase chain reaction to detect extracellular mRNAs. *Methods Mol. Biol.* **2011**, *784*, 15−25.
(82) Rani, S.; Clynes, M.; O'Driscoll, L. Detection of amplifiable mRNA extracellular to insulin-producing cells: potential for predicting beta cell mass and function. *Clin. Chem.* **2007**, *53*, 1936−1944.
(83) Takahashi, A.; Rahim, A.; Takeuchi, M.; Fukui, E.; Yoshizawa, M.; Mukai, K.; Suematsu, M.; Hasuwa, H.; Okabe, M.; Matsumoto, H. Impaired female fertility in tubulointerstitial antigen-like 1-deficient mice. *J. Reprod. Dev.* **2016**, *62*, 43−49.
(84) Carlsen, A. L.; Joergensen, M. T.; Knudsen, S.; de Muckadell, O. B. S.; Heegaard, N. H. H. Cell-free plasma microRNA in pancreatic ductal adenocarcinoma and disease controls. *Pancreas* **2013**, *42*, 1107−1113.
(85) Liu, P. F.; Jiang, W. H.; Han, Y. T.; He, L. F.; Zhang, H. L.; Ren, H. Integrated microRNA-mRNA analysis of pancreatic ductal adenocarcinoma. *Genet. Mol. Res.* **2015**, *14*, 10288−10297.
(86) Cnop, M.; Abdulkarim, B.; Bottu, G.; Cunha, D. A.; Igoillo-Esteve, M.; Masini, M.; Turatsinze, J. V.; Griebel, T.; Villate, O.; Santin, I.; Bugliani, M.; Ladriere, L.; Marselli, L.; McCarthy, M. I.; Marchetti, P.; Sammeth, M.; Eizirik, D. L. RNA sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. *Diabetes* **2014**, *63*, 1978−1993.
(87) Gebhardt, A.; Ackermann, W.; Ünver, N.; Elsässer, H. P. Expression of galectin-3 in the rat pancreas during regeneration following hormone-induced pancreatitis. *Cell Tissue Res.* **2004**, *315*, 321−329.
(88) Wang, L.; Friess, H.; Zhu, Z.; Frigeri, L.; Zimmermann, A.; Korc, M.; Berberat, P. O.; Büchler, M. W. Galectin-1 and galectin-3 in chronic pancreatitis. *Lab. Invest.* **2000**, *80*, 1233−1241.
(89) Muraro, M. J.; Dharmadhikari, G.; Grün, D.; Groen, N.; Dielen, T.; Jansen, E.; van Gurp, L.; Engelse, M. A.; Carlotti, F.; de Koning, E. J.; van Oudenaarden, A. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **2016**, *3*, 385−394.
(90) Baron, M.; Veres, A.; Wolock, S. L.; Faust, A. L.; Gaujoux, R.; Vetere, A.; Ryu, J. H.; Wagner, B. K.; Shen-Orr, S. S.; Klein, A. M.; Melton, D. A.; Yanai, I. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **2016**, *3*, 346−360 e4 .
(91) Segerstolpe, Å.; Palasantza, A.; Eliasson, P.; Andersson, E.-M.; Andréasson, A.-C.; Sun, X.; Picelli, S.; Sabirsh, A.; Clausen, M.; Bjursell, M. K.; Smith, D. M.; Kasper, M.; Ammala, C.; Sandberg, R. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **2016**, *24*, 593−607.
(92) Lawlor, N.; George, J.; Bolisetty, M.; Kursawe, R.; Sun, L.; Sivakamasundari, V.; Kycia, I.; Robson, P.; Stitzel, M. L. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **2017**, *27*, 208−222.
(93) Chen, L.; Li, J.; Zhang, Y.-H.; Feng, K.; Wang, S.; Zhang, Y.; Huang, T.; Kong, X.; Cai, Y.-D. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell Biochem.* **2018**, *119*, 3394−3403.
(94) Zhang, Y.-H.; Hu, Y.; Zhang, Y.; Hu, L.-D.; Kong, X. Distinguishing three subtypes of hematopoietic cells based on gene expression profiles using a support vector machine. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2018**, *1864*, 2255−2265.
(95) Chen, L.; Pan, X.; Zhang, Y.-H.; Kong, X.; Huang, T.; Cai, Y.-D. Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* **2018**, 0.
(96) Pan, X.; Hu, X.; Zhang, Y.; Feng, K.; Wang, S.; Chen, L.; Huang, T.; Cai, Y. Identifying Patients with Atrioventricular Septal Defect in Down Syndrome Populations by Using Self-Normalizing Neural Networks and Feature Selection. *Genes* **2018**, *9*, 208.
(97) Wang, S.; Cai, Y. Identification of the functional alteration signatures across different cancer types with support vector machine and feature analysis. *Biochim. Biophys. Acta, Mol. Basis Dis.* **2018**, *1864*, 2218−2227.
(98) Wang, D.; Li, J.-R.; Zhang, Y.-H.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of Differentially Expressed Genes between Original Breast Cancer and Xenograft Using Machine Learning Algorithms. *Genes* **2018**, *9*, 155.
(99) Dabrowski, M. J.; Draminski, M.; Diamanti, K.; Stepniak, K.; Mozolewska, M. A.; Teisseyre, P.; Koronacki, J.; Komorowski, J.; Kaminska, B.; Wojtas, B. Unveiling new interdependencies between significant DNA methylation sites, gene expression profiles and glioma patients survival. *Sci. Rep.* **2018**, *8*, 4390.
(100) Cui, S.; Youn, E.; Lee, J.; Maas, S. J. An Improved Systematic Approach to Predicting Transcription Factor Target Genes Using Support Vector Machine. *PLoS One* **2014**, *9*, No. e94519.
(101) Chen, L.; Zhang, S.; Pan, X.; Hu, X.; Zhang, Y.-H.; Yuan, F.; Huang, T.; Cai, Y.-D. HIV infection alters the human epigenetic landscape. *Gene Ther.* **2018**, 1.
(102) Pan, X.-Y.; Shen, H.-B. Robust Prediction of B-Factor Profile from Sequence Using Two-Stage SVR Based on Random Forest Feature Selection. *Protein Pept. Lett.* **2009**, *16*, 1447−1454.
(103) Mirza, A. H.; Berthelsen, C. H. B.; Seemann, S. E.; Pan, X.; Frederiksen, K. S.; Vilien, M.; Gorodkin, J.; Pociot, F. Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome Med.* **2015**, *7*, 39.
(104) Chen, L.; Pan, X.; Hu, X.; Zhang, Y.-H.; Wang, S.; Huang, T.; Cai, Y.-D. Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* **2018**, *143*, 1731−1740.
(105) Chen, L.; Wang, S.; Zhang, Y.-H.; Li, J.; Xing, Z.-H.; Yang, J.; Huang, T.; Cai, Y.-D. Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* **2017**, *5*, 26582−26590.
(106) Fang, Y.; Chen, L. A binary classifier for prediction of the types of metabolic pathway of chemicals. *Comb. Chem. High Throughput Screening* **2017**, *20*, 140−146.
(107) Guo, Z.-H.; Chen, L.; Zhao, X. A network integration method for deciphering the types of metabolic pathway of chemicals with heterogeneous information. *Comb. Chem. High Throughput Screening* **2018**, 670−680.
(108) Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics* **2004**, *20*, 2479−2481.
(109) Platt, J. *Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines*. Technical Report MSR-TR-98-14 1998.
(110) Dramiński, M.; Kierczak, M.; Nowak-Brzezińska, A.; Koronecki, J.; Komorowski, J. The Monte Carlo feature selection and interdependency discovery is unbiased. *Control Cybern.* **2011**, *40*, 199−211.
(111) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint*

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435

Conference on Artificial Intelligence; Lawrence Erlbaum Associates Inc., 1995; pp 1137-1145.
(112) Chen, L.; Zhang, Y.-H.; Huang, G.; Pan, X.; Wang, S.; Huang, T.; Cai, Y. D. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genomics* **2018**, *293*, 137−149.
(113) Ni, Q.; Chen, L. A feature and algorithm selection method for improving the prediction of protein structural class. *Comb. Chem. High Throughput Screening* **2017**, *20*, 612−621.
(114) Li, J.; Lu, L.; Zhang, Y.-H.; Liu, M.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* **2019**, *120*, 405−416.
(115) Zhao, X.; Chen, L.; Lu, J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* **2018**, *306*, 136−144.
(116) Chen, L.; Zhang, Y.-H.; Pan, X.; Liu, M.; Wang, S.; Huang, T.; Cai, Y.-D. Tissue Expression Difference between mRNAs and lncRNAs. *Int. J. Mol. Sci.* **2018**, *19*, 3416.
(117) Chen, L.; Zeng, W.-M.; Cai, Y.-D.; Feng, K.-Y.; Chou, K.-C. Predicting Anatomical Therapeutic Chemical (ATC) Classification of Drugs by Integrating Chemical-Chemical Interactions and Similarities. *PLoS One* **2012**, *7*, No. e35254.
(118) Chen, L.; Chu, C.; Zhang, Y.-H.; Zheng, M.; Zhu, L.; Kong, X.; Huang, T. Identification of Drug-Drug Interactions Using Chemical Interactions. *Curr. Bioinf.* **2017**, *12*, 526−534.
(119) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442−451.
(120) Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367−374.

DOI: 10.1021/acsomega.8b02171
ACS Omega 2019, 4, 6421−6435